Book Reviews

Ambiguity Resolution in Language Learning: Computational and Cognitive Models

Hinrich Schütze
(Xerox Palo Alto Research Center)

Stanford: CSLI Publications (Lecture notes number 71), 1997, xv+214 pp; distributed by Cambridge University Press; hardbound, ISBN 1-57586-075-9, $64.95; paperbound, ISBN 1-57586-074-0, $22.95

Reviewed by
I. Dan Melamed and Hang Li
West Group and NEC Corporation

1. Introduction

One of the proclaimed virtues of cognitive science as a field of study is that tremendous insights can be gained by investigating one of its component disciplines using tools and methods from another. In practice, this is not as easy as it appears, because the models employed in the investigation must account for complex cognitive phenomena. Yet, this was the goal that Schütze set for himself in this revised version of his 1995 dissertation. His ambition was to shed light on human language acquisition from computer science, and to present novel computational models for this human activity.

The book proposes computational models of the way people learn syntactic word categories, semantic word categories, and verb subcategorization frames. For the two word categorization tasks, the models are based on linear algebra over vector spaces. A neural network is employed to model acquisition of verb subcategorization frames. Schütze's models were the first to combine three important properties: the ability to learn from ambiguous input, gradient representations, and independence from external resources.

These properties are indispensable for a cognitive model of learning. Perception of ambiguous language is ubiquitous during human language learning, and people appear to have little difficulty in resolving ambiguities. Gradient representations allow Schütze's models to represent varying degrees of plausibility of competing internal linguistic forms. The author argues, from the success of his models in disambiguation during learning, that theories of human language acquisition should involve more gradient representations and fewer discrete ones. Perhaps the most attractive feature of Schütze's models from an engineering point of view is that they do not depend on any hand-built (external) resources other than corpus data. This property enables the models to be quickly adapted to new languages and new domains of application. Yet, Schütze's motivation for introducing this property had as much to do with cognitive-scientific theory as with engineering: He wanted to show that much can be learned without innate linguistic knowledge, thus challenging some of the received wisdom in linguistics.

The core text of the book is in three chapters, each dedicated to one of the three language acquisition problems mentioned above. A more detailed look into each of these chapters follows.

2. Syntactic Word Categorization

Chapter 2 focuses less on engineering issues and more on how syntactic word categorization can be accomplished by a cognitively plausible model. Its techniques are relatively simple by today's standards, but they are sufficient to support the author's claims on this topic.

An original feature of Schütze's syntactic categories is that they are not based on any external resource, for example, a predefined set of part-of-speech tags. Instead, they are induced from the input data by unsupervised clustering. The clustering takes place in TAG SPACE, a vector space where words are represented as feature vectors. The similarity of any two words is measured by comparing the cosine of their feature vectors, and clusters are formed on the basis of this similarity measure. The author intentionally restricts his features to those that can be automatically extracted from unannotated corpus data. Schütze uses singular value decomposition (SVD) to reduce the dimensionality of the vector space, and thus also the computational complexity of his clustering algorithms.[1] SVD also has the effect of filtering out some of the noise and ambiguity in the input data.

Although more sophisticated clustering techniques are known (e.g., Rose, Gurewitz, and Fox 1990), those described in the book were sufficient to demonstrate how *graded* syntactic categorization can be learned from rather impoverished training data. After defending the use of evaluation methods from the information retrieval literature, the author describes experiments with feature vectors that are based on progressively more contextual information. The results are quite impressive, given the simplicity of the method and the restriction to features that are visible in the unannotated corpus data.

3. Semantic Word Categorization

Much of the book's Chapter 3 on semantic word categorization is devoted to challenging traditional notions of word sense. The author argues from linguistics, psychology, and philosophy for his extension of Miller and Charles's (1991) "theory of senses as *groups of similar contexts*." Consequences of this theory include gradient representations of word senses, and coactivation of senses of a word in a particular context.

The chapter goes on to present WORD SPACE, a computational model of how people might learn semantic word categories so that their linguistic apparatus exhibits the properties predicted by the theory. WORD SPACE is mathematically similar to TAG SPACE, but it is inhabited by three different kinds of vectors: word vectors, context vectors, and sense vectors. Again, linear algebra and SVD are used to compute these vectors from nothing but corpus data.

The author presents a series of experiments, with pseudowords and with manually disambiguated ambiguous words, to demonstrate how well semantic categorization can be accomplished even with relatively simple mathematics and impoverished train-

1 The mathematics behind SVD and the clustering algorithms are reviewed in the book's appendix.

ing data. His results strongly undermine the notion that this task is impossible without innate linguistic ability. The chapter is also interesting from an engineering point of view. In contrast to syntactic categorization, there are several practical applications of semantic categorization of the kind where the categories need to be induced from the input data. Examples include the author's own work on information retrieval (Schütze and Pedersen 1995) as well as work on statistical translation models (Melamed 1998).

4. Verb Subcategorization Frames

Chapter 4 considers the problem of acquiring knowledge about whether a verb representing an action with a theme (e.g., a book) and a beneficiary (e.g., the student) allows the dative construction and/or the prepositional construction with *to*, as shown in the author's examples on page 121.

> She gave the student a book.
>
> She gave a book to the student.
>
> *She donated the church a book.
>
> She donated a book to the church.
>
> ... because He envied him the tree of life ...
>
> *He envied the tree of life to him.

The author's literature review on this topic concentrates on Lexical Rule Theory (LRT) (Pinker 1989), which is considered the most complete account of subcategorization acquisition. LRT assumes that different possible constructions of verbs are due to their different underlying morphological and lexico-semantic structures, represented by *discrete* features. In particular, it assumes that morphologically and lexico-semantically similar verbs form a class and members within the same class have the same or similar possible constructions. Thus the acquisition of subcategorization frames can be roughly viewed as a process of learning verb classes. The author points out several shortcomings of LRT. First, there are frequent exceptions in verb classes, and discrete representations cannot deal with this kind of uncertainty. Second, not only morphological and lexico-semantic features but also contextual features are useful in verb class learning.

To overcome these deficiencies, Schütze proposes a four-layer neural network, trained using the standard backpropagation algorithm. The training data is fed to the network one sentence at a time, where each sentence is represented by a set of morphological, lexico-semantic, and contextual features, as well as a feature representing the verb used in the sentence. For each sentence, the subcategorization frame (dative or prepositional construction) is also given. The model's output on a new sentence is a prediction about the most likely permissible constructions for the verb in the sentence. The large variety of features and the connectionist learning mechanism enable the model to make accurate subcategorization predictions not only for regular verbs but also for exceptional verbs.

5. Conclusions

The book is an enlightening investigation into the relationship between ambiguity resolution and language learning. It is unique in its focus on the disambiguation problem in learning, rather than in processing, in contrast to the main concern of much research in the NLP literature. Schütze's models are useful for some practical NLP tasks. For example, WORD SPACE can resolve word sense ambiguity and learn semantic classes solely from corpus data. This property makes the model attractive where word sense ambiguities are not resolved in training data and where it is inappropriate to rely on a prespecified semantic categorization. The models are also successful from the viewpoint of cognitive science, because they can account for some language acquisition phenomena that could not be explained by previous proposals.

References

Melamed, I. Dan. 1998. *Empirical Methods for Exploiting Parallel Texts*, Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6:1–28.

Pinker, Steven A. 1989. *Learnability and Cognition*, The MIT Press, Cambridge, MA.

Rose, Kenneth, E. Gurewitz and G. C. Fox. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters* 65(8):945–948.

Schütze, Hinrich and Jan O. Pedersen. 1995. Information retrieval based on word senses. *Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, 161–175.

I. Dan Melamed is a research scientist at West Group. His research interests span the field of empirical methods in natural language processing, with an emphasis on multilingual applications. Melamed's address is: West Group, 610 Opperman Drive, Eagan, MN, U.S.A., 55408; e-mail: melamed@research.westlaw.com; URL: http://www.cis.upenn.edu/~melamed/

Hang Li graduated from Kyoto University and received his Ph.D. from the University of Tokyo. He is currently a researcher at NEC Corporation. His technical interests include semantic knowledge acquisition and disambiguation in parsing. Li's address is: C&C Media Research Laboratories, NEC, 4-1-1 Miyazaki Miyamae-ku Kawasaki, 216-8555, Japan; e-mail: lihang@ccm.cl.nec.co.jp